Biophysical Journal
Article

Biophysical Society

Contact Potential for Structure Prediction of Proteins and Protein Complexes from Potts Model

Ivan Anishchenko,[1] Petras J. Kundrotas,[1,*] and Ilya A. Vakser[1,*]
[1]Computational Biology Program and Department of Molecular Biosciences, The University of Kansas, Lawrence, Kansas

ABSTRACT The energy function is the key component of protein modeling methodology. This work presents a semianalytical approach to the development of contact potentials for protein structure modeling. Residue-residue and atom-atom contact energies were derived by maximizing the probability of observing native sequences in a nonredundant set of protein structures. The optimization task was formulated as an inverse statistical mechanics problem applied to the Potts model. Its solution by pseudolikelihood maximization provides consistent estimates of coupling constants at atomic and residue levels. The best performance was achieved when interacting atoms were grouped according to their physicochemical properties. For individual protein structures, the performance of the contact potentials in distinguishing near-native structures from the decoys is similar to the top-performing scoring functions. The potentials also yielded significant improvement in the protein docking success rates. The potentials recapitulated experimentally determined protein stability changes upon point mutations and protein-protein binding affinities. The approach offers a different perspective on knowledge-based potentials and may serve as the basis for their further development.

INTRODUCTION

Computer simulations are essential for studying biological macromolecules, including proteins. Along with the search space sampling, the energy function is the key component of modeling. The energy function can be either derived from the general physical principles (like a number of popular force fields (1–4)) or based on diverse sets of known protein structures (various knowledge-based or statistical potentials (5–8)). Statistical potentials provide the balance between accuracy and computational efficiency. Thus, they are successfully applied to many problems, such as discrimination of the native structure from decoys (9,10), fold recognition (11), structure prediction (12), protein docking (13–15) and design (16,17), and prediction of protein stability and affinity (18–20). Simplified energy models provide insight into general principles of protein folding and binding (21–24).

One of the common approaches to developing statistical potentials is to calculate the probability of various structural features observed in a set of experimental protein structures relative to a reference state (7,25). The probability is subsequently converted into energy using the inverse Boltzmann relation (7). However, the choice of the reference state, which serves as an imaginary protein model without interactions, is not a well-defined problem, and a number of approximations have been proposed. Among them are averaging (26), finite ideal-gas (10), spherical noninteracting (27), atom-shuffled (28), random-walk chain (29), and quasichemical (6) approximations. A different strategy for deriving statistical potentials is based on optimization of the energy parameters to maximize recognition of the native structure from a set of decoys (30–32). Despite the success of statistical potentials in various applications, their physical interpretation is not quite clear (33–35). Thus, derivation of the potential that provides fundamental and transparent insight is highly desirable.

Many problems that require describing direct (microscopic) interactions of objects (atoms, particles, etc.) from observation of microscopic configurations of the system of these objects can be successfully tackled by inverse statistical mechanics approaches (see (36–38) and references therein). In particular, the Ising (39) and Potts (40) models were used to study the collective behavior of neurons (41,42), infer gene-interaction networks from experimentally observed transcription profiles (43), predict residue-residue contacts from multiple sequence alignments

Submitted March 7, 2018, and accepted for publication July 31, 2018.

*Correspondence: pkundro@ku.edu or vakser@ku.edu

Ivan Anishchenko's present address is Department of Biochemistry, University of Washington, Seattle, Washington.

Editor: Amedeo Caflisch.

https://doi.org/10.1016/j.bpj.2018.07.035

(37,44,45), study protein fitness landscapes (46,47), and infer epistatic effects from fitness (48). Interaction parameters in these models are often recovered using the maximal entropy principle (49), resulting in the least structured (i.e., most generic) model that is still consistent with the experimental data. In this study, we show that inverse statistical mechanics formalism applied to the Potts model can be used to construct both residue-residue and atom-atom contact potentials, with the latter outperforming most existing energy functions in a number of tests. A closely related approach has already been utilized to derive residue-residue statistical contact potentials (50–52). However, these studies have not gained much attention, most likely because they were discussed from a different perspective, i.e., protein evolution and design, and no detailed analysis of the performance of the constructed potentials in protein structural modeling has been reported. In this article, we bridge this gap and show that the inverse Potts inference can be applied to construct simple but effective residue-residue and atom-atom contact potentials, with the latter performing on par with the best existing statistical potentials. The effectiveness of the potentials is attributed to 1) the consistent estimation of the energy parameters by the pseudolikelihood maximization approach and 2) explicit treatment of one-body energies at the learning stage.

MATERIALS AND METHODS

Contact energies

Graph representation of protein structure

Noncovalent interactions in a protein can be modeled by a simple contact potential, suggesting that if two structural elements (called here interaction centers) are closer in space than a cutoff distance $d_{\max}$, then these elements contribute some distance-independent value to the total energy. The interaction center can be the center of mass of a residue (for residue-residue potentials) or a single heavy atom (for atom-atom potentials). The number of distinct types of interaction centers (hereafter denoted as q) can vary depending on the level of generalization. In this study, we consider one residue-residue (RRCE20) and three atom-atom (AACE18, AACE20, and AACE167) contact potentials (RRCE and AACE are residue-residue and atom-atom contact energies, respectively; the summary description of the potentials is in Table 1). In the AACE18 potential, the atoms are grouped according to their physicochemical properties (19), yielding $q = 18$ distinct atom types. For the AACE20 potential, all heavy atoms in a residue are grouped together, resulting in $q = 20$ atom types. In the most detailed AACE167 potential, each heavy atom in the 20 residue types is considered separately, yielding $q = 167$ atom types. Hydrogen atoms or different protonation states of titratable amino acids were not considered.

For the applications discussed in this work, it is sufficient to represent a single protein structure by an undirected graph $G_p(V_p,E_p)$ (Fig. 1). In such a graph, the set of nodes $V_p = \{v_i\}$ includes all interaction centers for a protein p. The set of edges $E_p = \{e_{ij}\}$ comprises connections between interaction centers i and j, which are 1) closer in space than a cutoff distance $d_{\max}$ and 2) separated by at least $k_{\min}$ residues in the protein sequence (Fig. 1 *A*). For a given protein, the number of nodes L is fixed, but the number of edges may vary with the protein conformation. The only free parameters are $d_{\max}$ and $k_{\min}$. Their optimal values are to be determined by the benchmarking. Besides $k_{\min}$, there are no other assumptions on the protein topology (e.g., information on the intraresidue connectivity is not used).

Energy of protein

Each node in the graph $G_p(V_p,E_p)$ can adopt one of q possible states (q is determined solely by the type of the potential; in this study, $q = 18$, 20, or 167; see Table 1). For clarity, a state of graph $G_p(V_p,E_p)$ is denoted by the same letters as the graph vertices $\{v_i\}$, giving a vector

$$\vec{\mathbf{v}} = (v_1, \ldots, v_i, \ldots, v_L), \tag{1}$$

which is composed of integer numbers $v_i \in (1,...,k,...,q)$, associated with atom (residue) types of the graph nodes. To derive the contact potentials, we introduce one- and two-body energy terms to account for self-energies and energies of contacting atom (residue) pairs within the protein. Thus, each graph node i of type v_i can be associated with one of q possible numbers h_{v_i} from vector

$$\vec{\mathbf{h}} = (h_1,\ldots, h_k, \ldots, h_q). \tag{2}$$

In turn, each edge, e_{ij}, can be attributed to one of $q \times q$ values J_{v_i,v_j} from a symmetric matrix

$$\mathbf{J} = \begin{pmatrix} J_{11} & \cdots & J_{1k} & \cdots & J_{1q} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ J_{k1} & \cdots & J_{kk} & \cdots & J_{kq} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ J_{q1} & \cdots & J_{qk} & \cdots & J_{qq} \end{pmatrix}, \tag{3}$$

depending on types v_i and v_j of nodes i and j, respectively. For every type of potential, there are unique sets of $\vec{\mathbf{h}}$ and $\mathbf{J}$ parameters shared by all nodes and edges of the graph. However, each protein has a unique graph $G_p(V_p,E_p)$ and atom (residue) type assignment vector $\vec{\mathbf{v}}$, which are solely determined by the conformation and amino acid composition of that protein.

Summation over the nodes and edges of the graph $G_p(V_p,E_p)$ yields an expression for the energy of the protein

$$U(\vec{\mathbf{v}}) = \sum_{\{v_i\}} h_{v_i} + \sum_{\{e_{ij}\}} J_{v_i,v_j}. \tag{4}$$

It is similar to the expression for the energy (Hamiltonian) of the q-state generalized Potts model in statistical physics (40), in which pairwise interactions depend on the states of the interacting sites and the local fields (or self-energies) act on the single sites of the system.

TABLE 1 Four Types Of Contact Potentials

Potential	Interaction centers	Number of Interaction Center Types, q	Description of Types	Number of Parameters
RRCE20	residue centroids	20	20 standard amino acids	230
AACE18	heavy atoms	18	18 atom types from (19)	189
AACE20	heavy atoms	20	20 standard amino acids	230
AACE167	heavy atoms	167	all heavy atoms in 20 standard amino acids	14,195

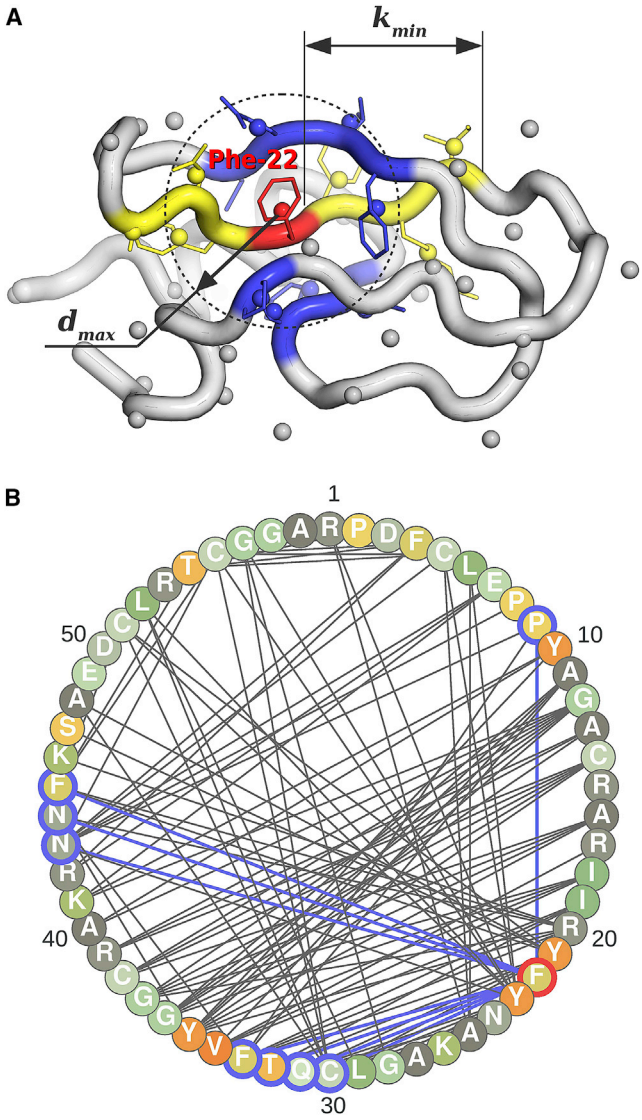

FIGURE 1 Graphical model of protein 3D structure. (*A*) A cartoon representation of the 58-residue bovine pancreatic trypsin inhibitor mutant 1G6X is shown. Residue centroids are shown by small spheres. Blue parts are an example of residue neighborhood, showing all residues with centroids within $d_{\max} = 7$ Å from Phe22 (in *red*). Residues in yellow are within $k_{\min} = 3$ positions in sequence from Phe22 and are not included in its neighborhood when calculating the parameters of the potential. (*B*) The graph $G_p(V_p,E_p)$ is a simplified representation of the BPTI mutant structure at $d_{\max} = 7$ Å and $k_{\min} = 3$. Phe22 and its neighbors have red and blue borders, respectively. Nodes of the graph are color-coded according to the amino acid type to indicate their state.

For a fixed graph $G_p(V_p,E_p)$ (i.e., fixed protein conformation), the probability of state $\vec{\mathbf{v}}$ is given by the Boltzmann (Gibbs) distribution

$$P(\vec{\mathbf{v}}) = \frac{1}{Z}\exp(-\beta U(\vec{\mathbf{v}})), \tag{5}$$

where β is a scaling factor (which, in statistical physics, means the inverse energy of thermal fluctuations at temperature T, $\beta = 1/RT$), and Z is the statistical sum over the set of all possible system states $\{\vec{\mathbf{v}}\}$:

$$Z = \sum_{\{\vec{\mathbf{v}}\}}\exp(-\beta U(\vec{\mathbf{v}})). \tag{6}$$

Parameters $\vec{\mathbf{h}}$ and $\mathbf{J}$ (Eqs. 2 and 3) are not known a priori but can be inferred from a large set of known protein structures (see below). Probability distribution in Eq. 5 is also known as Markov random field (53) on graph $G_p(V_p,E_p)$.

Pseudolikelihood approximation

Because native protein sequences are close to optimal for their three-dimensional structures (54), for a given structure of a protein, an accurate energy model should assign highest probability to the native sequence compared to any non-native one. This concept, for example, helps in protein design when one tries to find a sequence that best fits a given protein fold (55). In terms of the energy function (Eq. 4), the task can be formulated as an optimization problem of finding values of $\vec{\mathbf{h}}$ (Eq. 2) and $\mathbf{J}$ (Eq. 3) that maximize the probability of observing the native state

$$P(\vec{\mathbf{v}}_{\text{nat}}) = \max_{\{\vec{\mathbf{v}}\}}(P(\vec{\mathbf{v}})). \tag{7}$$

However, the optimization problem (Eq. 7) cannot be solved directly because of the combinatorial complexity of the partition function (Eq. 6). To make the problem tractable, the probability function (Eqs. 5 and 6) for the native state (sequence) is approximated by a product of local conditional probabilities (pseudolikelihoods)

$$P(\vec{\mathbf{v}}_{\text{nat}}) \approx \prod_i \frac{\exp(-\beta U(v_{i,\text{nat}}))}{\sum_{k=1}^{q}\exp(-\beta U(v_{i,k}))}, \tag{8}$$

where multiplication is performed over all atoms (residues) and summation in the denominator is over all possible q states of a single interaction center. $U(v_{i,k})$ is the "energy" of a single interaction center, or $G_p(V_p,E_p)$ node, in state k

$$U(v_{i,k}) = h_k + \sum_{\{e_{ij}\}} J_{kk'}, \tag{9}$$

where summation is performed over all other nodes in $G_p(V_p,E_p)$ connected to node i by an edge. The temperature factor $\beta = 1$ is used throughout the work. In the pseudolikelihood approximation, all nodes are in the native states, and only the state of a current node varies to calculate the "pseudo"-statistical sum (denominator in Eq. 8). The pseudolikelihoods are known to provide asymptotically consistent estimates of parameters $\vec{\mathbf{h}}$ and $\mathbf{J}$ (56) and are successfully applied to large sample size problems in physics and biology (37,51,57).

In the above formalism, the optimization problem (Eq. 7) is reduced to solving the system of differential equations

$$\begin{cases} \dfrac{\partial(-\log P(\vec{\mathbf{v}}_{\text{nat}}))}{\partial h_k} = 0, \quad k = 1,\ldots,q \\ \vdots \\ \dfrac{\partial(-\log P(\vec{\mathbf{v}}_{\text{nat}}))}{\partial J_{kk'}} = 0, \quad k = 1,\ldots,q \text{ and } k' = 1,\ldots,q \end{cases}. \tag{10}$$

For convenience, in the analytical deduction of the derivatives in Eq. 10, we used negative pseudo-log-likelihoods. More details of the pseudolikelihood optimization are in the Supporting Materials and Methods.

The solution to system of equations (Eq. 10) within the graph $G_p(V_p,E_p)$ would provide $\vec{\mathbf{h}}$ and $\mathbf{J}$ specific only for one protein. To obtain generic potentials, we solved the system of equations (Eq. 10) for a composite graph $G(V,E)$ constructed by joining graphs $G_p(V_p,E_p)$ for all individual proteins in a large set of protein structures (details on this "training" set are in Materials and Methods). The product in Eq. 8 runs over all nodes in that composite graph, whereas all other considerations remain the same.

Once $\vec{\mathbf{h}}$ and $\mathbf{J}$ values are obtained for the training set, only the $\mathbf{J}$ matrix, which constitutes the contact potential, is applied to several problems in protein modeling (see Results). The role of self-energies $\vec{\mathbf{h}}$ is to provide an accurate estimation of $\mathbf{J}$ (also discussed in the Results).

It is worth noting that in terms of the graph representation, the development of "classical" statistical potentials is usually limited to collecting information on the number of nodes in the composite graph in state k and the number of edges connecting nodes in states k and k' without solving Eq. 10.

Training set of protein structures

To calculate the Potts parameters $\vec{\mathbf{h}}$ and $\mathbf{J}$, a nonredundant training set of 6338 protein chains was collected by the Protein Sequence Culling Server (58). Only x-ray structures with resolution $\leq$2.0 Å, R-factor $\leq$0.25, and $\geq$40 residues per chain were selected. Redundancy was removed at 25% sequence identity cutoff. Individual chains were extracted from Protein Data Bank (PDB) asymmetric units, and missing heavy atoms were restored by the PDB2PQR software (59) using CHARMM topology parameters (1). Alternative residue conformations, if present in the original PDB structure, were removed by the same PDB2PQR program. Nineteen chains from the initial pool of 6338 structures could not be processed either because of multiple models in the original PDB file or a large number of missing heavy atoms (>10%). These structures were left out from the consideration, yielding the final set of 6319 chains from 6092 different PDB entries.

Parameters of the contact potential

At each value of the distance cutoff $d_{\max}$ (from 4 to 15 Å, with 0.1 Å step) and sequence separation $k_{\min}$ (from 1 to 10, with step 1), we built the graph $G(V,E)$ for 6319 single protein structures from the training set. Minimization of the objective function (Eq. 8) with derivatives (Eq. 10) was performed by the in-house C program specifically designed for this purpose. The GNU Scientific Library (http://www.gnu.org/software/gsl/) implementation of the quasi-Newton Broyden-Fletcher-Goldfarb-Shanno method (60) (bgfs2 module of the GNU Scientific Library) was used. Minimization started with all the target parameters set to zero and proceeded iteratively until the norm of the gradient achieved the absolute tolerance of 10^{-3}.

CASP decoys

Decoys for near-native structure detection were compiled from all tertiary structure predictions submitted to Critical Assessment of Structure Prediction (CASP) rounds X and XI (61,62). Following the CASP practice, the models were analyzed at the level of evaluation units, or domains, assigned by the assessors. To make the energy estimates consistent, partial models with incomplete chains were removed from the pool of decoys. Overall, 224 domains from 172 protein chains were selected for testing. PDB files of models and the tables with models' parameters and ranking according to global distance test, total score (GDT_TS) (63) were downloaded from the CASP repository (http://predictioncenter.org/download_area/).

Scoring of CASP decoys

Performance of different energy functions on the CASP decoys were assessed in terms of the Z-score, defined as the distance (measured in standard deviations) between the energy of the best (highest GDT_TS score) model U_b (or the native structure) calculated by the tested function and the mean energy of all decoy structures $\langle U\rangle$

$$\text{Z-score} = \frac{U_b - \langle U\rangle}{\sigma}, \tag{11}$$

where σ is the standard deviation of energy U (given by the tested function) for all decoys in the set. In addition to the Z-score (Eq. 11), we also used the Pearson correlation coefficient between energies and GDT_TS scores of the models, as well as the normalized rank of the best-energy model

$$1 - R = 1 - \frac{R_{\text{GDT\_TS}}}{N_{\text{tot}}}, \tag{12}$$

where $R_{\text{GDT\_TS}}$ is the rank by the GDT_TS score of the best-energy model, and N_{tot} is the total number of the models in the set. The form $1 - R$ was used to transform the normalized rank to the increasing function of the scoring method effectiveness.

For the assessment of the potentials, the energy of a protein model was calculated by simple summation of the inferred couplings $\mathbf{J}$ over all pairs (i,j) of interaction centers that are consistent with distance cutoff $d_{\max}$ used to derive the potential

$$U_p = \sum_{(i,j)} J_{v_i,v_j}. \tag{13}$$

The sum over the self-energies (local fields) $\vec{\mathbf{h}}$ was omitted in Eq. 13 because it does not depend on the protein conformation and thus does not affect ranking of the model structures. Subscripts v_i, v_j enumerate types of atoms (residues) i and j, respectively ($v_i \in 1,2,...,q$, where q = 18, 20, or 167 depending on the contact potential type; see Table 1).

Data set of point mutations

To evaluate applicability of the contact potentials to prediction of the change in the folding free energy $\Delta\Delta G$ upon mutations, we used a data set of 2648 point mutations for 131 globular proteins with experimentally resolved x-ray or NMR structure, for which mutation-induced change in protein stability was determined experimentally (64). The set is derived from the ProTherm database (65). 235 mutations in the set originate from NMR structures (12 distinct PDB IDs) with the number of models from 5 to 46. Because the data set contains experimentally resolved structures for the wild-type proteins (those deposited in the PDB) only, the structure of a mutant was obtained by manual replacement of the corresponding side chain. The replacement was followed by the SCWRL4 (66) repacking of the residues within 6 Å distance to the mutated residue (the residue-residue distance defined as any atom to any atom of the two residues). To compensate for possible biases introduced by SCWRL4, the same relaxation procedure was applied to the same residues of the wild-type structure. For each mutated residue X, relative solvent exposure was calculated as the ratio between the absolute solvent-accessible surface area (SASA) of this residue in the wild-type structure and the reference SASA for this type of residue in the Gly-X-Gly tripeptide (67). A residue was considered to be at the surface if >20% of its SASA was exposed.

Assuming that the folding free energy of a protein is proportional to its internal energy in the folded state, $\Delta\Delta G_{\text{calc}}$ was approximated by the difference in internal energies (Eq. 13) of the mutant and the wild-type:

$$\Delta\Delta G_{\text{calc}} = U_{\text{mut}} - U_{\text{WT}}. \tag{14}$$

Docking decoys

The initial set of 1020 binary protein-protein complexes, for which the structures of the complex (in PDB biological assembly) and the structure of both unbound components are available, was generated by the ProPairs tool (68) run locally on a PDB snapshot with the default parameters. The set was postprocessed to retain only pairs with a high similarity of bound and unbound partners (sequence identity >96% and coverage >80% (69)), which reduced the set size to 427. Additional purging of structurally similar (template modeling-score (70) >0.8) and large (>2500 residues per interactor) complexes yielded the final set of 396 complexes (DOCKGROUND Benchmark 4.0 (71) http://dockground.compbio.ku.edu). For comparison, we also used 230 protein-protein complexes in the docking Benchmark 5.0 from Weng's group (69).

The unbound proteins from both benchmarks were docked by the fast Fourier transform rigid-body docking program GRAMM (72,73) at low resolution, with 3.5 Å grid step and 10° angular interval. The top 100,000 matches per complex, ranked solely by the shape complementarity, were compared to the reference complex obtained by structural superposition of the unbound monomers onto corresponding proteins in the co-crystallized complex. The quality of the docking models was assessed by the Critical Assessment of Predicted Interactions (CAPRI) criteria (74) (Table S1). Docking success rate was defined as the fraction of complexes for which at least one successful prediction (defined at different accuracy categories) was in the top n predictions. The docking predictions were further reranked by the energy of the proteins A and B interface U_{AB}, calculated (similarly as for individual proteins; see Eq. 13) by summing up the couplings $\mathbf{J}$ over all pairs ($i \in$ A, $j \in$ B) of the interchain contacts closer in space than d_{max}:

$$U_{AB} = \sum_{(i \in A, j \in B)} J_{v_i, v_j}. \quad (15)$$

The predicted matches were clustered to identify the most probable hit within each putative docking funnel. Only one lowest energy prediction from each cluster was selected.

Affinity benchmark

To access how U_{AB} (Eq. 15) correlates with the protein-protein binding affinities, the set of 92 protein-protein complexes with known co-crystallized structures and experimentally determined binding affinities ΔG_{exp} was selected from the Affinity Benchmark version 2.0 (69). We considered only the rigid-body cases—those without significant conformational changes upon binding. Such cases were defined as bound/unbound interface root mean-square distance < 1.5 Å and fraction of non-native contacts (the number of non-native residue-residue contacts in the predicted complex divided by the total number of contacts in that complex (74)) < 0.4.

Energy functions

For comparison, we tested the following knowledge-based energy functions: discrete optimized protein energy (DOPE) (27), distance-scaled, finite ideal-gas reference (DFIRE) (10), dipolar DFIRE (dDFIRE) (75,76), random walk (RW) and RWplus (29), generalized orientation-dependent all-atom potential (GOAP) (77), OPUS-PSP (78), RF-HA-SRS (28), and RF-CB-SRS-OD (79). DOPE, DFIRE, RW, and RF-HA-SRS are all-heavy-atom distance-dependent potentials, whereas RWplus, dDFIRE, and GOAP have an additional orientation-dependent term. OPUS-PSP is an orientation-dependent contact potential defined for blocks of side-chain atoms. RF-CB-SRS-OD is a residue-level distance- and orientation-dependent energy function. In addition, a simple residue-residue contact potential by Miyazawa and Jernigan MJ3h (80) was also tested because of its best performance in scoring of protein docking decoys (81,82).

RESULTS AND DISCUSSION

Parameters of the contact potentials

Different distance cutoffs d_{max} and sequence separations k_{min} may result in a different graph model for the protein structure (Fig. 1) and thus in different sets of local fields $\vec{\mathbf{h}}$ and couplings $\mathbf{J}$ that maximize the likelihood function (Eq. 8). To find the optimal d_{max} and k_{min} values, we derived our four potentials RRCE20, AACE20, AACE167, and AACE18 (Table 1) using 1110 various d_{max} and k_{min} combinations (see Materials and Methods). The performance of the potentials was evaluated by discriminating best models from the CASP decoys (Figs. 2 and S1).

All four potentials performed poorly when contacts between residues adjacent in the sequence ($k_{min} \gtrsim 1$) were considered in the derivation of the contact energies. Residues that are close in sequence are close in space primarily because of the covalent bonds. Thus, taking such contacts into account obscures the treatment of nonbonded interactions, especially at smaller d_{max} (83). As d_{max} increases, more interacting pairs contribute to the potentials, and the

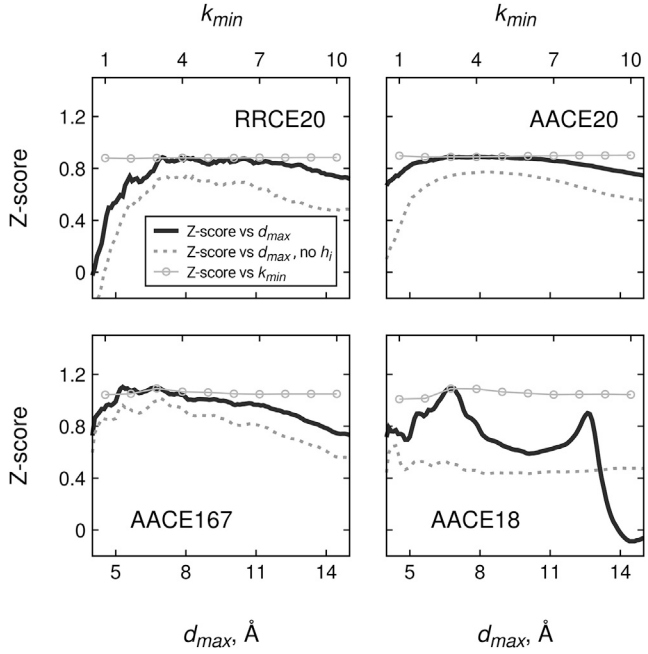

FIGURE 2 Performance of residue-residue and atom-atom contact potentials in best-model structure recognition from CASP decoys. Statistical potentials derived at different values of sequence separation k_{min} and distance cutoff d_{max} were used to score models of 224 protein domains submitted to CASP rounds X and XI. Performance is measured as Z-score of the highest GDT_TS score model averaged over all 224 evaluation units. The solid line shows distance dependence of the Z-scores (*bottom horizontal axis*) at the optimal sequence separation $k_{min} = 3$. Performance of the potentials derived without local fields $\vec{\mathbf{h}}$ is shown by the dashed lines. Thin gray lines with circles show the Z-score dependence on the sequence separation k_{min} (*top horizontal axis*) at $d_{max} = 8.0$ Å (RRCE20, AACE20) and $d_{max} = 6.9$ Å (AACE167, AACE18). The plots are cross-sections of the heat maps in Fig. S1 at specific values of sequence separation k_{min} and distance cutoff d_{max}.

relative contribution of sequence-adjacent residues declines rapidly. However, with complete exclusion of the sequence-adjacent residues, some portion of the nonbonded interaction energy remains unaccounted for, causing a drop, albeit slight, in the performance. The optimal performance was observed at $k_{\min} = 3$. Thus, the potentials derived at this $k_{\min}$ value were used in the further analysis unless stated otherwise. Despite high correlation of the RRCE20 contact energies and the well-known Miyazawa-Jernigan matrix MJ3h (80) ($R = 0.90$), the former still shows better decoy discrimination by all three measures (Fig. 4).

The RRCE20 and AACE20 potentials showed very similar trends with varying $d_{\max}$. This suggested that if all atoms within one residue are assigned to one type, the way of calculating contacts (either between residue centroids or between residue heavy atoms) has a negligible effect on the energy function. The optimal performance for these potentials was achieved within a broad $d_{\max}$ interval (6–11 Å). If the protein heavy atoms were split into 167 different types, the trend remained similar. However, the best performance is observed at lower $d_{\max} = 5 \div 8$ Å; Z-score increases significantly from 0.89 to 0.88 for RRCE20 and AACE20, respectively, to 1.09 for AACE167.

The distinct feature of the AACE18 potential is the two sharp peaks of enhanced performance (in terms of Z-score, quantitatively similar to the performance of the much more complex AACE167) at $d_{\max} \sim 6.9$ and ~12.6 Å (Fig. 2 *D*). The exact reason for the two-peak distribution is not clear because the statistical potential parameters are derived from a self-consistent solution of the system of equations (10), in which elucidating the effect of a particular factor is nontrivial if possible at all. However, the peaks do not appear when the potential is derived using a reduced model (leaving out h_{v_i} terms in Eqs. 4 and 9; *dashed lines* in Fig. 2 *D*). This points to an important interconnectivity between one- and two-body energies, which substantially elevates the efficiency of the potential at certain $d_{\max}$. Generally, the reduced model yields significantly less effective contact potentials (the *dashed lines* in Fig. 2 are all below the *solid lines*). AACE18 correlates poorly ($R = 0.43$) with the original atomic contact energies from (19).

The convergence analysis of the inferred energy parameters showed that they are already close to optimal. Thus, further increase in the number of structures deposited to PDB can only marginally improve the potentials (Figs. S2 and S3).

Local fields

After the local fields $\vec{\mathbf{h}}$ and couplings $\mathbf{J}$ are learned by solving the system of equations (Eq. 10), one-body terms (or self-energies) can be omitted in protein structure energy estimates and scoring applications (Eqs. 13, 14, and 15). Nevertheless, the local fields are essential internal parameters of the model at the learning stage (Eqs. 7, 8, and 9), boosting the effectiveness of the resulting two-body energies $\mathbf{J}$ (*dashed* and *solid lines* in Fig. 2). Below, we provide a detailed analysis of how the trained h_k parameters are related to the basic features of individual interaction center types.

Empirically, we found that the one-body energies $\vec{\mathbf{h}}$ can be described by a linear combination of two interaction centers features (Fig. 3 *B*, *rightmost panel*)

$$h_k = a \times p_k + b \times n_k + c, \tag{16}$$

where p_k and n_k are the propensity (frequency of occurrence) and the average coordination number (in graph terminology, the average number of node neighbors in the graph $G(V,E)$ connected by an edge) for the interaction centers of type k in the training set, respectively. The coordination number is inversely related to the exposure of the interaction center to the solvent because the interface atoms or residues have less contact with other interaction centers than those in the protein core.

The two parameters p_k and n_k contribute $\geq 90\%$ to the fields $\vec{\mathbf{h}}$ for all potentials and $d_{\max}$ (Fig. 3, *A–D*). For the RRCE20 and AACE20 potentials (Fig. 3, *A* and *B*), the propensities contribute significantly more at smaller distances, whereas at larger $d_{\max}$, the coordination numbers become more important. The AACE167 potential showed qualitatively similar behavior. The main difference was that contribution of the propensities became larger at significantly smaller $d_{\max}$ (a steep *dark-gray peak* on the *left-hand side* of Fig. 3 *A*). The AACE18 potential showed distinctly different patterns (Fig. 3 *A*, *rightmost panel*). For this potential, the relative importance of p_k and n_k weakly depends on $d_{\max}$, and the propensities generally have a higher contribution to local fields than the coordination numbers (e.g., at optimal $d_{\max} = 6.9$ Å, n_k-values contribute ~65 and 36% to the local fields for the AACE167 and AACE18 potentials, respectively).

Discrimination of protein near-native structures

The quality of knowledge-based energy functions is often assessed by their ability to recognize the native structure or the best model in a set of decoys (9,84,85). Models submitted to the CASP competition (86) are believed to be the most challenging (87) and have been recently used by others to benchmark their statistical potentials (79,88). Thus, we tested our four potentials on the CASP decoys from rounds X and XI of the competition (61,62). Identifying the best model from the decoys (that also corresponds to the real-case modeling scenario, when the native structure is not known) is generally more challenging than identifying the native structure (79,87). This is also the case for our potentials: Z-scores for the best model are on average in the 0.9–1.1 range, whereas corresponding numbers for the

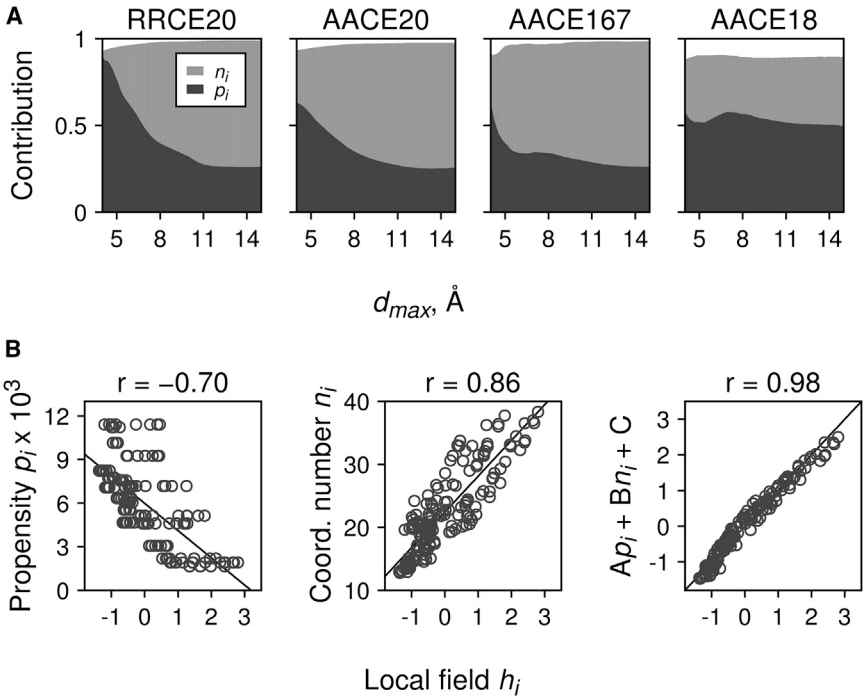

FIGURE 3 Properties of the local fields $\vec{\mathbf{h}}$. (*A*) The contribution (relative importance) of atom propensities p_k and average coordination numbers n_k to local fields $\vec{\mathbf{h}}$ for the least-squares linear model (Eq. 16) with varying cutoff distance $d_{\max}$ is shown, calculated for RRCE20, AACE20, AACE167, and AACE18 potentials, respectively. The data was obtained by the averaging-over-orderings method by Lindeman et al. (94) as implemented in the *relaimpo* package (95) for R statistical computing language. (*B*) As an example, h_k correlations with atom propensities p_k and coordination numbers n_k and their linear combination (Eq. 16) are shown for the AACE167 potential derived at $d_{\max} = 6.9$ Å and $k_{\min} = 3$.

native structure are 0.6–0.8 Z-score units higher (Fig. S4). Although discrimination of the best model is significantly harder than that of the native structure (which also substantially complicates comparison of different scoring functions), it is more relevant to the real-case scenario when only the models, but not the native structure, are available. Thus, unlike a number of other studies on scoring functions, we focused our analysis on the ability of the potentials to discriminate the best model and compared their performance to 10 state-of-the-art knowledge-based energy functions (Fig. 4).

In terms of the Z-score (Eq. 11), the AACE18 and AACE167 potentials are among the top ones, only behind GOAP (77). Assessment by the normalized rank (Eq. 12) also puts AACE18 and AACE167 on top of the list ($1 - R = 0.809$ and 0.808, respectively), only slightly behind GOAP ($1 - R = 0.821$). High correlation of energy and structural accuracy scores of the decoys indicates good scoring (77). In this respect, AACE18 shows the best performance (the average correlation coefficient 0.606), followed by GOAP (0.587) and AACE167 (0.585) (Fig. 4). These values are similar to correlations reported elsewhere (e.g., (77)). Statistical analysis reveals, however, that the differences between AACE18, AACE167, and GOAP are marginal and all three potentials have comparable performance, significantly better than the other tested energy functions (Table S2). The other two assessment scores (Z-score and normalized rank) are not as discriminative (see corresponding *p*-values in Table S2). However, they still place AACE18 and AACE167 among the top ones, only slightly behind GOAP. The way to establish exact ranking for the 14 potentials is not obvious. However, consistency between the three assessment scores should indicate that GOAP, AACE18, and AACE167 are the three best-performing energy functions in the best-model discrimination test. In discrimination of the native structure test (Fig. 4 *D*), the performance order is slightly different, with OPUS-PSP and RF-HA-SRS on top of the list by the Z-score, followed by GOAP and AACE167. This suggests that some energy functions are more tuned for high-resolution decoys but are less successful in discriminating models of moderate accuracy. In this respect, AACE167 is more sensitive in selecting the native structure compared to AACE18 (Fig. 4 *D*).

The residue-level RRCE20 and AACE20 potentials perform poorly by all measures, suggesting the need for atomic details for effective contact potentials. On the other hand, our contact potentials AACE18 and AACE167, which are quite simple (e.g., no distance or orientation dependency), are sufficient for capturing most structural details of the protein models, which usually is achieved by much more complex energy functions.

Protein stability changes upon point mutations

The top-performing AACE18 and AACE167 potentials were further tested for their ability to predict the change in protein stability $\Delta\Delta G$ upon single mutation (see Materials and Methods). The testing was done on the benchmark set of 2648 point mutations (64) in terms of Pearson's r for correlation of the calculated $\Delta\Delta G_{\text{calc}}$ (Eq. 14) and experimental $\Delta\Delta G_{\text{exp}}$, separately for the buried and the exposed

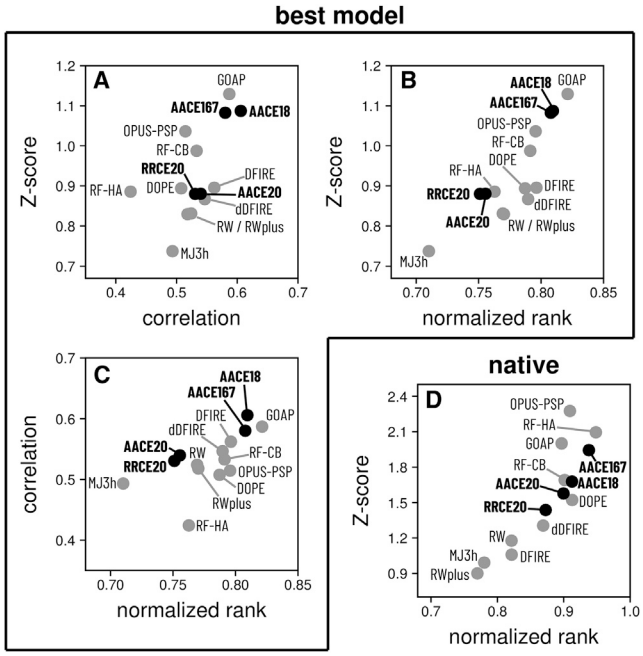

FIGURE 4 Performance of various energy functions in the best-model structure recognition from CASP decoys. The best model's Z-score, its normalized rank $1 - R$, and Pearson's correlation coefficient r of the energy score and GDT_TS score of models, all averaged over 224 CASP decoy sets, are shown for different scoring functions on scatter plots (*A*)–(*C*) (one plot per each combination of the above three assessment scores). For comparison, average Z-scores and normalized ranks for the native structure are shown on plot (*D*).

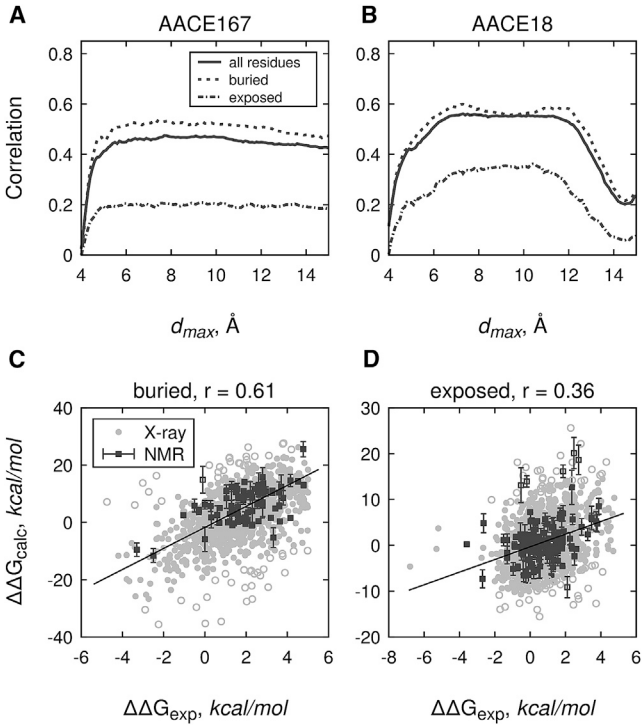

FIGURE 5 Prediction of protein stability changes upon point mutations by the AACE167 and AACE18 potentials. Experimentally determined $\Delta\Delta G$ values for 2648 point mutations from 131 proteins are correlated with the ones calculated by AACE167 (*A*) and AACE18 (*B*) potentials at different cutoff distances d_{max}. As an example, correlations for the AACE18 potential at $d_{max} = 6.9$ Å and $k_{min} = 3$ are shown separately for (*C*) buried (relative SASA ≤ 0.2, 1429 residues) and (*D*) exposed residues (relative SASA > 0.2, 1219 residues), respectively. Light gray circles correspond to the x-ray structures. Dark gray squares are based on multimodel NMR structures and show calculated $\Delta\Delta G$ values averaged over all states, with the error bars showing standard deviations. All points with deviations from the least-squares linear fit (*solid black lines*) not falling into (0.025, 0.975) percentile range were treated as outliers, shown by open circles/squares.

residues at various distance cutoffs d_{max} (Fig. 5). Correlations were calculated after removal of outliers, which include all points with deviations from the least-squares linear fit outside a 2.5–97.5% range. Surface residues are generally more susceptible to structural variations because their side chains are less constrained by the neighbors. In addition, there is a significant solvent contribution to their energetics. These effects are especially hard to account for by any energy function. Indeed, both potentials had a significant drop in performance (by ~50% in terms of r) for the surface residues compared to the performance for the buried residues (Fig. 5, *A* and *B*). Overall, the AACE167 performance almost saturates at $r \sim 0.45$ for $d_{max} > 5$ Å. However, the predictions for the surface residues are much less accurate compared to the predictions for the buried ones ($r \sim 0.2$ and 0.5, respectively). For $d_{max} < 5$ Å, the performance drops almost to zero regardless of the residue exposure to solvent.

The AACE18 energy function also has generally better predictions for the buried residues than for the exposed ones. However, the performance is not constant at $d_{max} > 5$ Å (Fig. 5 *B*). Similar to the best-model recognition from the CASP decoys (Fig. 2 *D*), the elevated r values are observed for the buried residues at two d_{max} values, 7 and 12 Å. For the surface residues, however, such peaks are not observed, and the best recapitulation of the experimental energies is achieved at $d_{max} \sim 8 \div 11$ Å (Fig. 5 *B*). Interestingly, this d_{max} region coincides with the region of lower performance on the buried residues. This distance range roughly corresponds to the water-mediated interactions (89), which indicates that solvent effects are better treated by the simpler AACE18 rather than by the more complicated AACE167 potential.

An example of correlation between $\Delta\Delta G_{exp}$ and $\Delta\Delta G_{calc}$ calculated by the AACE18 potential for buried and exposed residues (Fig. 5, *C* and *D*) indicates that NMR structures yield slightly more accurate $\Delta\Delta G_{calc}$ estimates than the x-ray structures ($r = 0.65$ vs. 0.58 for buried and 0.38 vs. 0.34 for exposed residues, correspondingly). This might be related to a more adequate environment of the NMR models and to averaging over the ensemble of all models in the PDB entry. However, a direct comparison is problematic because the sets of NMR and x-ray structures consist of different proteins.

In comparison with other energy functions, AACE18 potential is ranked sixth, with $r_{AACE18} = 0.554$ compared to $r_{dDFIRE} = 0.591$ of the top performing dDFIRE (Fig. S6 *A*, one-sided p-value = 0.023 at 95% confidence). However, the difference in correlations between the first five energy functions is not statistically significant (p-value = 0.189 between the first—dDFIRE—and the fifth—DOPE). The performance of the other three potentials, AACE167, AACE20, and RRCE20, is significantly worse (Fig. S6 *A*).

Contact potentials in protein docking

Predicting protein-protein complexes from the structures of the individual monomers (protein docking) remains a challenging problem in computational structural biology because of a fine balance between different factors (shape complementarity, solvent and electrostatic effects, conformational changes, etc.) that enable specific binding but are hard to accurately account for. Thus, the protein docking problem is often addressed by coarse-grained approaches, at least at the initial modeling stages (90).

We tested how well our contact potentials score the low-resolution docking predictions by GRAMM (72,73) (Figs. 6 and S5). Models of complexes were assessed according to the CAPRI criteria (Table S1). Predictions with acceptable and better quality were considered successful. Similar to the CASP decoys (Fig. S1), the best discrimination of the near-native models is achieved at the distance cutoffs $d_{max} = 6.9$ Å for AACE167 and AACE18 potentials and $d_{max} = 8.0$ Å for RRCE20 and AACE20. However, the best performance is achieved at larger sequence separation $k_{min} = 5$. All four energy functions, in most cases, outperform the Miyazawa-Jernigan MJ3h statistical potential (80), which has been recently shown to be one of the top-performing scoring functions in protein docking (81). The largest improvement over MJ3h is achieved by the atom contact potentials AACE167 and AACE18. The AACE18 also proved its efficiency in discriminating near-native docking matches in a recent joint CASP/CAPRI round of the CASP12 competition (91).

Interestingly, the reference structure often has a higher (worse) energy score than the top near-native docking clusters (Fig. S7). This is likely caused by atom clashes in the reference structure obtained by simple structural superimposition of the unbound monomers onto corresponding bound conformations. Our docking protocol, albeit low resolution, is able to find better-scoring near-native matches. The true native conformation is generally scored higher in the case of AACE18 and in particular AACE167 potentials (Fig. S7), suggesting that taking into account protein flexibility might further improve the ranking. However, as Fig. S7 shows, selection of the native conformation from the docking decoys is still difficult. Similar success rates were reported previously for popular protein-protein docking scoring functions ZRANK (92) and integration of residue- and atom-based potentials for docking (93).

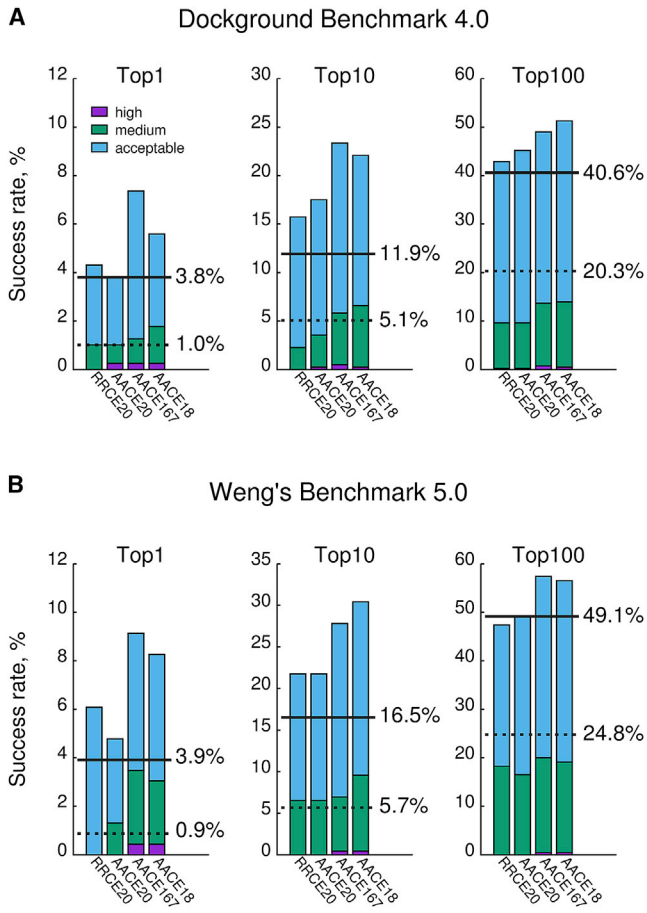

FIGURE 6 Scoring of low-resolution docking decoys. The top 100,000 matches per complex with highest shape complementarity score from GRAMM were evaluated by RRCE20, AACE20, AACE167, and AACE18 contact potentials, followed by L-root-mean-square-deviation-based clustering with 10 Å radius, for (*A*) Dockground Benchmark 4 and (*B*) Weng's Benchmark 5. The lowest energy model from each cluster was further assessed by the CAPRI criteria (see Table S1). The docking success rate for 1, 10, and 100 best scored clusters was calculated (*bars*). Dashed lines are the baselines of the success rates when models are ranked according to the raw shape complementarity. Solid lines are docking success rates attained by the Miyazawa-Jernigan MJ3h statistical potential.

Correlation with protein binding affinity

Finally, we analyzed correlation of the interchain energy U_{AB} (Eq. 15) or ΔG_{calc} calculated by the atom AACE18 and AACE167 potentials at various distance cutoffs d_{max} for the protein complexes in the affinity benchmark (69), with the experimentally determined binding affinities ΔG_{exp} (Figs. 7 and S6 *B*). The experimental binding free energies were recapitulated significantly worse by the more complex AACE167 potential than by the simpler AACE18 (Fig. 7 *A*). Even a naïve ΔG predictor, which approximates binding

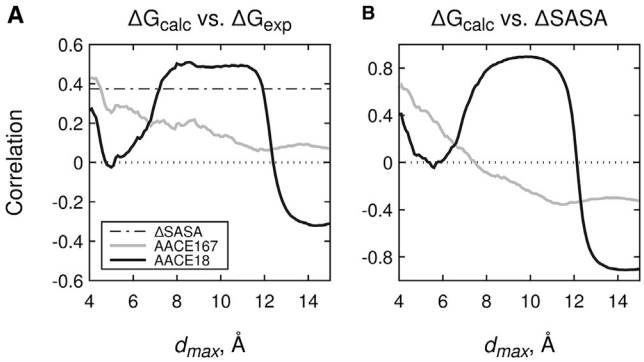

FIGURE 7 Prediction of proteins binding affinities by AACE167 and AACE18 potentials. (*A*) Experimentally determined binding free energies (ΔG_{exp}) for 92 rigid-body complexes from affinity benchmark 2 are correlated with the ones calculated (ΔG_{calc}) by the AACE167 and AACE18 potentials with varying cutoff distances d_{max}. Dashed line at 0.375 show performance of a naïve predictor, which approximates binding free energy by the change in solvent-accessible surface area (ΔSASA) upon complex formation. (*B*) The correlation of calculated binding free energies ΔG_{calc} and ΔSASA values is shown.

free energy by the change in the SASA (ΔSASA) upon complex formation, outperformed AACE167 at almost all d_{max} (except small $d_{max} \sim 4$ Å). At the same time, AACE18 again performed significantly better at $d_{max} = 8 \div 11$ Å than at other distances, which correlates with the data on the point mutations for the exposed residues (Fig. 5 *B*). In this d_{max} range, the AACE18 energies tend to be highly correlated with ΔSASA (Fig. 7 *B*), which is not the case for the more complex AACE167. This indicates that desolvation effects are largely captured by the AACE18 potential, albeit in a simple form $\Delta G_{desolvation} \sim \Delta$SASA. However, AACE18 performance is still superior to the naïve ΔSASA predictor (Fig. 7 *A*) as well as all other energy functions tested on the affinity benchmark ($r_{AACE18} = 0.508$, followed by the DFIRE potential with $r_{DFIRE} = 0.445$; Fig. S6 *B*). In comparison, specialized affinity prediction algorithms still have a better performance, with correlations up to $r = 0.53$ for the full set and $r = 0.75$ for rigid-body cases (69).

CONCLUSIONS

In summary, we presented a framework for generating semi-empirical general-purpose contact potentials for proteins structure modeling. The potentials are derived from the Potts model by solving the inverse statistical physics problem. The model contains only two adjustable parameters, interaction distance cutoff d_{max} and separation in the sequence for the interacting units (residues or atoms) k_{min}. No other assumptions on the protein topology or information on intraresidue connectivity were used. Unlike many other statistical potentials, our derivation scheme explicitly includes one-body energy terms, which are shown to be a significant component of the model, boosting the effectiveness of the derived potentials.

The potentials were derived purely from the structural data in the PDB and are completely independent of any reference state. The results showed that they are successful not only in recognizing near-native models of individual proteins but also in scoring of protein docking decoys, recapitulating the experimental binding energies, and predicting stability changes upon point mutations. Such transferability of atomic potentials is strongly dependent on the assignment of the atom types. Among three considered assignment schemes, the grouping of atoms according to their physicochemical properties yielded consistently top-performing AACE18 potential. Interestingly, despite the effectiveness of the most detailed AACE167 potential in scoring of CASP decoys, it was much less effective at recapitulating experimental free energies. Large number of atom types enables fine-tuning of the AACE167 potential to achieve high decoy discrimination rate. However, at the same time, it affects its transferability to other applications.

It should be also noted that even for the most effective AACE18 potential, it is hard, if possible at all, to use one optimal contact distance d_{max} for all applications. For example, for discrimination of the structural decoys (for both the individual proteins and the protein complexes), the optimal d_{max} value was 6.9 Å. However, $d_{max} = 8.0$ Å yielded better correlation with the experimentally determined binding free energies. This discrepancy was shown to be at least partially related to the solvent effects, which are not explicitly taken into account by our model.

Overall, it is quite remarkable that in a wide range of protein structure modeling applications, simple contact potentials with no distance or orientation dependencies are sufficient for the same or better performance than much more complex knowledge-based energy functions used in the field. However, such simplicity may also pose limitations to the potentials applicability, e.g., in structure refinement, because of the lack of sensitivity to small changes in atom-atom distances inherent to the contact potentials. Complementing contact potentials with other scoring terms (e.g., the extent of clashes, surface area, etc.) is one way to overcome this problem, which has been explored by us in CASP-CAPRI competition (91).

In the future, we plan further development of the statistical potentials by incorporating distance dependence and solvent effects as well as exploring higher-order interactions (e.g., including three-body terms in Eq. 4). We will also explore different atom types, including hydrogen atoms and different protonation states of the titratable residues. On the learning side, more thorough selection of the training set, as well as inclusion of the interchain contacts from biological assemblies in PDB, could also lead to better contact energy estimates. All these questions can be addressed within the approach presented in this work.

The potentials are available at http://vakser.compbio.ku.edu/main/resources.php

SUPPORTING MATERIAL

Supporting Materials and Methods, seven figures, and two tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(18)30923-8.

AUTHOR CONTRIBUTIONS

I.A. was responsible for concept, design, acquisition, analysis and interpretation of data, and writing of the manuscript. P.J.K. was responsible for supervision, analysis and interpretation of data, and writing of the article. I.A.V. was responsible for supervision, analysis and interpretation of data, and writing of the article. All authors read and approved the final manuscript.

ACKNOWLEDGMENTS

This study was supported by National Institutes of Health grant R01GM074255 and National Science Foundation grants DBI1262621 and DBI1565107.

REFERENCES

1. MacKerell, A. D., D. Bashford, …, M. Karplus. 1998. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B.* 102:3586–3616.
2. Cornell, W. D., P. Cieplak, …, P. A. Kollman. 1996. A second generation force field for the simulation of proteins, nucleic acids, and organic molecules. *J. Am. Chem. Soc.* 118:2309–2309.
3. Jorgensen, W. L., D. S. Maxwell, and J. Tirado-Rives. 1996. Development and testing of the OPLS all-atom force field on conformational energetics and properties of organic liquids. *J. Am. Chem. Soc.* 118:11225–11236.
4. Oostenbrink, C., A. Villa, …, W. F. van Gunsteren. 2004. A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6. *J. Comput. Chem.* 25:1656–1676.
5. Tanaka, S., and H. A. Scheraga. 1976. Medium- and long-range interaction parameters between amino acids for predicting three-dimensional structures of proteins. *Macromolecules.* 9:945–950.
6. Miyazawa, S., and R. L. Jernigan. 1985. Estimation of effective interresidue contact energies from protein crystal structures: quasi-chemical approximation. *Macromolecules.* 18:534–552.
7. Sippl, M. J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* 213:859–883.
8. Lazaridis, T., and M. Karplus. 2000. Effective energy functions for protein structure prediction. *Curr. Opin. Struct. Biol.* 10:139–145.
9. Park, B., and M. Levitt. 1996. Energy functions that discriminate X-ray and near native folds from well-constructed decoys. *J. Mol. Biol.* 258:367–392.
10. Zhou, H., and Y. Zhou. 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Sci.* 11:2714–2726.
11. Buchete, N. V., J. E. Straub, and D. Thirumalai. 2004. Development of novel statistical potentials for protein fold recognition. *Curr. Opin. Struct. Biol.* 14:225–232.
12. Skolnick, J. 2006. In quest of an empirical potential for protein structure prediction. *Curr. Opin. Struct. Biol.* 16:166–171.
13. Mintseris, J., and Z. Weng. 2003. Atomic contact vectors in protein-protein recognition. *Proteins.* 53:629–639.
14. Chuang, G. Y., D. Kozakov, …, S. Vajda. 2008. DARS (decoys as the reference state) potentials for protein-protein docking. *Biophys. J.* 95:4217–4227.
15. Liu, S., and I. A. Vakser. 2011. DECK: distance and environment-dependent, coarse-grained, knowledge-based potentials for protein-protein docking. *BMC Bioinformatics.* 12:280.
16. Hu, C., X. Li, and J. Liang. 2004. Developing optimal non-linear scoring function for protein design. *Bioinformatics.* 20:3080–3098.
17. Boas, F. E., and P. B. Harbury. 2007. Potential energy functions for protein design. *Curr. Opin. Struct. Biol.* 17:199–204.
18. Guerois, R., J. E. Nielsen, and L. Serrano. 2002. Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320:369–387.
19. Zhang, C., G. Vasmatzis, …, C. DeLisi. 1997. Determination of atomic desolvation energies from the structures of crystallized proteins. *J. Mol. Biol.* 267:707–726.
20. Bordner, A. J., and R. A. Abagyan. 2004. Large-scale prediction of protein geometry and stability changes for arbitrary single point mutations. *Proteins.* 57:400–413.
21. Bryngelson, J. D., and P. G. Wolynes. 1987. Spin glasses and the statistical mechanics of protein folding. *Proc. Natl. Acad. Sci. USA.* 84:7524–7528.
22. Li, H., C. Tang, and N. S. Wingreen. 1997. Nature of driving force for protein folding: a result from analyzing the statistical potential. *Phys. Rev. Lett.* 79:765–768.
23. Mirny, L., and E. Shakhnovich. 2001. Protein folding theory: from lattice to all-atom models. *Annu. Rev. Biophys. Biomol. Struct.* 30:361–396.
24. Pokarowski, P., A. Kloczkowski, …, A. Kolinski. 2005. Inferring ideal amino acid interaction forms from statistical protein contact potentials. *Proteins.* 59:49–57.
25. Betancourt, M. R., and D. Thirumalai. 1999. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* 8:361–369.
26. Samudrala, R., and J. Moult. 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *J. Mol. Biol.* 275:895–916.
27. Shen, M. Y., and A. Sali. 2006. Statistical potential for assessment and prediction of protein structures. *Protein Sci.* 15:2507–2524.
28. Rykunov, D., and A. Fiser. 2007. Effects of amino acid composition, finite size of proteins, and sparse statistics on distance-dependent statistical pair potentials. *Proteins.* 67:559–568.
29. Zhang, J., and Y. Zhang. 2010. A novel side-chain orientation dependent potential derived from random-walk reference state for protein fold selection and structure prediction. *PLoS One.* 5:e15386.
30. Goldstein, R. A., Z. A. Luthey-Schulten, and P. G. Wolynes. 1992. Protein tertiary structure recognition using optimized Hamiltonians with local interactions. *Proc. Natl. Acad. Sci. USA.* 89:9029–9033.
31. Maiorov, V. N., and G. M. Crippen. 1992. Contact potential that recognizes the correct folding of globular proteins. *J. Mol. Biol.* 227:876–888.
32. Thomas, P. D., and K. A. Dill. 1996. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA.* 93:11628–11633.
33. Thomas, P. D., and K. A. Dill. 1996. Statistical potentials extracted from protein structures: how accurate are they? *J. Mol. Biol.* 257:457–469.
34. BenNaim, A. 1997. Statistical potentials extracted from protein structures: are these meaningful potentials? *J. Chem. Phys.* 107:3698–3706.
35. Hamelryck, T., M. Borg, …, J. Ferkinghoff-Borg. 2010. Potentials of mean force for protein structure prediction vindicated, formalized and generalized. *PLoS One.* 5:e13714.
36. Habeck, M. 2014. Bayesian approach to inverse statistical mechanics. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 89:052113.

37. Ekeberg, M., C. Lövkvist, …, E. Aurell. 2013. Improved contact prediction in proteins: using pseudolikelihoods to infer Potts models. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 87:012707.

38. Levy, R. M., A. Haldane, and W. F. Flynn. 2017. Potts Hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Curr. Opin. Struct. Biol.* 43:55–62.

39. Brush, S. G. 1967. History of the Lenz-Ising model. *Rev. Mod. Phys.* 39:883–893.

40. Wu, F. Y. 1982. The Potts model. *Rev. Mod. Phys.* 54:235–268.

41. Schneidman, E., M. J. Berry, II, …, W. Bialek. 2006. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature.* 440:1007–1012.

42. Cocco, S., S. Leibler, and R. Monasson. 2009. Neuronal couplings between retinal ganglion cells inferred by efficient inverse statistical physics methods. *Proc. Natl. Acad. Sci. USA.* 106:14058–14062.

43. Lezon, T. R., J. R. Banavar, …, N. V. Fedoroff. 2006. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Natl. Acad. Sci. USA.* 103:19033–19038.

44. Marks, D. S., L. J. Colwell, …, C. Sander. 2011. Protein 3D structure computed from evolutionary sequence variation. *PLoS One.* 6:e28766.

45. Balakrishnan, S., H. Kamisetty, …, C. J. Langmead. 2011. Learning generative models for protein fold families. *Proteins.* 79:1061–1078.

46. Figliuzzi, M., H. Jacquier, …, M. Weigt. 2016. Coevolutionary landscape inference and the context-dependence of mutations in beta-lactamase TEM-1. *Mol. Biol. Evol.* 33:268–280.

47. Shekhar, K., C. F. Ruberman, …, A. K. Chakraborty. 2013. Spin models inferred from patient-derived viral sequence data faithfully describe HIV fitness landscapes. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.* 88:062705.

48. Flynn, W. F., A. Haldane, …, R. M. Levy. 2017. Inference of epistatic effects leading to entrenchment and drug resistance in HIV-1 protease. *Mol. Biol. Evol.* 34:1291–1306.

49. Jaynes, E. T. 1957. Information theory and statistical mechanics. *Phys. Rev.* 106:620–630.

50. Bonnard, C., C. L. Kleinman, …, N. Lartillot. 2009. Fast optimization of statistical potentials for structurally constrained phylogenetic models. *BMC Evol. Biol.* 9:227.

51. Kleinman, C. L., N. Rodrigue, …, N. Lartillot. 2006. A maximum likelihood framework for protein design. *BMC Bioinformatics.* 7:326.

52. Zhou, X., and S. C. Schmidler. 2009. Bayesian Parameter Estimation in Ising and Potts Models: A Comparative Study with Applications to Protein Modeling. Department of Statistical Science, Duke University, Durham, NC.

53. Kindermann, R., J. L. Snell; American Mathematical Society. 1980. Markov Random Fields and Their Applications. American Mathematical Society, Providence, RI.

54. Kuhlman, B., and D. Baker. 2000. Native protein sequences are close to optimal for their structures. *Proc. Natl. Acad. Sci. USA.* 97:10383–10388.

55. Kuhlman, B., G. Dantas, …, D. Baker. 2003. Design of a novel globular protein fold with atomic-level accuracy. *Science.* 302:1364–1368.

56. Gidas, B. 1988. Consistency of maximum likelihood and pseudo-likelihood estimators for Gibbs distributions. *In* Stochastic Differential Systems, Stochastic Control Theory and Applications. W. Fleming and P.-L. Lions, eds. Springer, pp. 129–145.

57. Aurell, E., and M. Ekeberg. 2012. Inverse ising inference using all the data. *Phys. Rev. Lett.* 108:090201.

58. Wang, G., and R. L. Dunbrack, Jr. 2005. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res.* 33:W94–W98.

59. Dolinsky, T. J., J. E. Nielsen, …, N. A. Baker. 2004. PDB2PQR: an automated pipeline for the setup of Poisson-Boltzmann electrostatics calculations. *Nucleic Acids Res.* 32:W665–W667.

60. Fletcher, R., and R. Fletcher. 2000. Structure of methods. Practical Methods of Optimization. John Wiley & Sons, Ltd., pp. 12–43.

61. Moult, J., K. Fidelis, …, A. Tramontano. 2016. Critical assessment of methods of protein structure prediction: progress and new directions in round XI. *Proteins.* 84 (*Suppl 1*):4–14.

62. Moult, J., K. Fidelis, …, A. Tramontano. 2014. Critical assessment of methods of protein structure prediction (CASP)–round x. *Proteins.* 82 (*Suppl 2*):1–6.

63. Zemla, A. 2003. LGA: a method for finding 3D similarities in protein structures. *Nucleic Acids Res.* 31:3370–3374.

64. Dehouck, Y., A. Grosfils, …, M. Rooman. 2009. Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0. *Bioinformatics.* 25:2537–2543.

65. Bava, K. A., M. M. Gromiha, …, A. Sarai. 2004. ProTherm, version 4.0: thermodynamic database for proteins and mutants. *Nucleic Acids Res.* 32:D120–D121.

66. Krivov, G. G., M. V. Shapovalov, and R. L. Dunbrack, Jr. 2009. Improved prediction of protein side-chain conformations with SCWRL4. *Proteins.* 77:778–795.

67. Rose, G. D., A. R. Geselowitz, …, M. H. Zehfus. 1985. Hydrophobicity of amino acid residues in globular proteins. *Science.* 229:834–838.

68. Krull, F., G. Korff, …, E. W. Knapp. 2015. ProPairs: a data set for protein-protein docking. *J. Chem. Inf. Model.* 55:1495–1507.

69. Vreven, T., I. H. Moal, …, Z. Weng. 2015. Updates to the integrated protein-protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. *J. Mol. Biol.* 427:3031–3041.

70. Zhang, Y., and J. Skolnick. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins.* 57:702–710.

71. Kundrotas, P. J., I. Anishchenko, …, I. A. Vakser. 2018. Dockground: a comprehensive data resource for modeling of protein complexes. *Protein Sci.* 27:172–181.

72. Katchalski-Katzir, E., I. Shariv, …, I. A. Vakser. 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. *Proc. Natl. Acad. Sci. USA.* 89:2195–2199.

73. Vakser, I. A. 1995. Protein docking for low-resolution structures. *Protein Eng.* 8:371–377.

74. Méndez, R., R. Leplae, …, S. J. Wodak. 2003. Assessment of blind predictions of protein-protein interactions: current status of docking methods. *Proteins.* 52:51–67.

75. Yang, Y., and Y. Zhou. 2008. Specific interactions for ab initio folding of protein terminal regions with secondary structures. *Proteins.* 72:793–803.

76. Yang, Y., and Y. Zhou. 2008. Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. *Protein Sci.* 17:1212–1219.

77. Zhou, H., and J. Skolnick. 2011. GOAP: a generalized orientation-dependent, all-atom statistical potential for protein structure prediction. *Biophys. J.* 101:2043–2052.

78. Lu, M., A. D. Dousis, and J. Ma. 2008. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. *J. Mol. Biol.* 376:288–301.

79. Rykunov, D., and A. Fiser. 2010. New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics.* 11:128.

80. Miyazawa, S., and R. L. Jernigan. 1999. Self-consistent estimation of inter-residue protein contact energies based on an equilibrium mixture approximation of residues. *Proteins.* 34:49–68.

81. Moal, I. H., M. Torchala, …, J. Fernández-Recio. 2013. The scoring of poses in protein-protein docking: current capabilities and future directions. *BMC Bioinformatics.* 14:286.

82. Anishchenko, I., P. J. Kundrotas, and I. A. Vakser. 2017. Modeling complexes of modeled proteins. *Proteins.* 85:470–478.

83. Melo, F., R. Sánchez, and A. Sali. 2002. Statistical potentials for fold assessment. *Protein Sci.* 11:430–448.

84. Samudrala, R., and M. Levitt. 2000. Decoys 'R' Us: a database of incorrect conformations to improve protein structure prediction. *Protein Sci.* 9:1399–1401.

85. Simons, K. T., C. Kooperberg, …, D. Baker. 1997. Assembly of protein tertiary structures from fragments with similar local sequences using simulated annealing and Bayesian scoring functions. *J. Mol. Biol.* 268:209–225.

86. Moult, J., J. T. Pedersen, …, K. Fidelis. 1995. A large-scale experiment to assess protein structure prediction methods. *Proteins.* 23:ii–v.

87. Handl, J., J. Knowles, and S. C. Lovell. 2009. Artefacts and biases affecting the evaluation of scoring functions on decoy sets for protein structure prediction. *Bioinformatics.* 25:1271–1279.

88. Cossio, P., D. Granata, …, A. Trovato. 2012. A simple and efficient statistical potential for scoring ensembles of protein structures. *Sci. Rep.* 2:351.

89. Levy, Y., and J. N. Onuchic. 2006. Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct.* 35:389–415.

90. Vakser, I. A. 2013. Low-resolution structural modeling of protein interactome. *Curr. Opin. Struct. Biol.* 23:198–205.

91. Kundrotas, P. J., I. Anishchenko, …, I. A. Vakser. 2018. Modeling CAPRI targets 110-120 by template-based and free docking using contact potential and combined scoring function. *Proteins.* 86 (*Suppl 1*):302–310.

92. Pierce, B., and Z. Weng. 2007. ZRANK: reranking protein docking predictions with an optimized energy function. *Proteins.* 67:1078–1086.

93. Vreven, T., H. Hwang, and Z. Weng. 2011. Integrating atom-based and residue-based scoring functions for protein-protein docking. *Protein Sci.* 20:1576–1586.

94. Lindeman, R. H., P. F. Merenda, and R. Z. Gold. 1980. Introduction to bivariate and multivariate analysis. Scott, Foresman and Comp, Glenview, IL.

95. Grömping, U. 2006. Relative importance for linear regression in R: the package relaimpo. *J. Stat. Softw.* 17:1–27.

Biophysical Journal, Volume *115*

Supplemental Information

Contact Potential for Structure Prediction of Proteins and Protein Complexes from Potts Model

Ivan Anishchenko, Petras J. Kundrotas, and Ilya A. Vakser

SUPPOPRTING MATERIAL

Finding energy parameters by pseudo-likelihood maximization

Substituting Eq. 8 into Eq. 7 in the main text gives the pseudo-likelihood function L_p of parameters $\vec{\mathbf{h}}$ and $\mathbf{J}$:

$$L_p\left(\vec{\mathbf{h}},\mathbf{J}\right)=\prod_{i=1}^{L}\frac{\exp\left(-\beta U\left(x_i^{\text{nat}};N_i^{\text{nat}}\right)\right)}{\sum_{m=1}^{q}\exp\left(-\beta U\left(m;N_i^{\text{nat}}\right)\right)}. \tag{A1}$$

$U\left(k;N_i\right)$ is the energy of a single interaction center in state *k* surrounded by the set of neighbors N_i, which includes all atoms (residues) connected to site *i* by an edge in the graph (shown in blue in Fig. 1A , B of the main text), and "nat" indicates that all the neighbors are in their native states. Atom (residue) types x_i^{nat} along with the neighbors N_i^{nat} are from the native structures of the proteins in the training set and are fixed throughout the computations. For computational efficiency, we convert the pseudo-likelihood in Eq. A1 to the negative pseudo-log-likelihood function, which transforms the optimization problem (Eq. 7 in the main text) to

$$-\log\left(L_p\right)=\sum_{i=1}^{L}\left[\beta U\left(x_i^{\text{nat}};N_i^{\text{nat}}\right)+\log\sum_{k=1}^{q}\exp\left(-\beta U\left(k;N_i^{\text{nat}}\right)\right)\right]\xrightarrow[\vec{\mathbf{h}},\mathbf{J}]{}\min \tag{A2}$$

The gradient of the negative pseudo-log-likelihood function has components

$$\begin{cases}\dfrac{\partial\left(-\log L_p\right)}{\partial h_a}=\beta\sum_{i=1}^{L}\left[\delta_{a,x_i^{\text{nat}}}-P_i\left(a\right)\right],\\[2ex]\dfrac{\partial\left(-\log L_p\right)}{\partial J_{aa}}=\beta\sum_{i=1}^{L}\left(\sum_{j\in N_i^{\text{nat}}}\delta_{a,x_j^{\text{nat}}}\right)\left[\delta_{a,x_i^{\text{nat}}}-P_i\left(a\right)\right],\\[2ex]\dfrac{\partial\left(-\log L_p\right)}{\partial J_{ab}}=\beta\sum_{i=1}^{L}\left(\sum_{j\in N_i^{\text{nat}}}\delta_{a,x_j^{\text{nat}}}\right)\left[\delta_{a,x_i^{\text{nat}}}-P_i\left(a\right)\right]+\beta\sum_{i=1}^{L}\left(\sum_{j\in N_i^{\text{nat}}}\delta_{b,x_j^{\text{nat}}}\right)\left[\delta_{b,x_i^{\text{nat}}}-P_i\left(b\right)\right],\quad a<b\end{cases} \tag{A3}$$

where $a,b=1,...,q$, $\delta_{a,b}$ is the Kronecker delta and

$$P_i(a)=\frac{\exp\left(-\beta U\left(a;N_i^{\text{nat}}\right)\right)}{\sum_{k=1}^{q}\exp\left(-\beta U\left(k;N_i^{\text{nat}}\right)\right)} \tag{A4}$$

is the conditional probability of observing site i in state a, provided all neighboring sites N_i^{nat} are in their native states. We explicitly force the coupling matrix **J** to be symmetric by aggregating off-diagonal contributions from J_{ab} and J_{ba} into one derivative (3rd line in Eq. A3) and ommiting the lower triangular part of **J** (i.e. $a>b$) from computations. This reduces the total number of unknowns to $q+q\cdot(q+1)/2$. Given analytic derivatives in Eq. A3, the optimization problem Eq. A2 can be efficiently solved (e.g. by a Quasi-Newton method), until the requirement $\nabla\left(-\log L_p\right)\simeq 0$ (Eq. A3) is met.

Table S1. Docking accuracy according to CAPRI criteria

Quality category	Condition
High	f_{nat}[(1)] ≥ 0.5 *and* (L-RMSD[(2)] ≤ 1.0 Å *or* I-RMSD[(3)] ≤ 1.0 Å)
Medium	f_{nat} ≥ 0.3 *and* (1.0 < L-RMSD ≤ 5.0 Å *or* 1.0 < I-RMSD ≤ 2.0 Å)
Acceptable	f_{nat} ≥ 0.1 *and* (5.0 < L-RMSD ≤ 10.0 Å *or* 2.0 < I-RMSD ≤ 4.0 Å)
Incorrect	f_{nat} < 0.1 *and* (L-RMSD > 10.0 Å *and* I-RMSD > 4.0 Å)

(1) Fraction of predicted native residue–residue contacts

(2) C^{α} ligand RMSD when receptors are optimally aligned

(3) Interface C^{α} RMSD calculated over the set of native interface residues after a structural superposition of these residues

Table S2. *Details of various energy functions performance in the best model recognition from CASP decoys*. Best model's Z-score, its normalized rank 1 – *R*, and Pearson's correlation coefficient *r* of the energy score and GDT_TS score of models, all averaged over 224 CASP decoy sets, are shown in columns 6, 9 and 2 respectively. 95% confidence interval for the correlation coefficient averaged over 224 decoys is in column 3. 14 energy functions were ordered according to their *r* values, and one- and two-sided Wilcoxon signed-rank test was applied to compare samples of 224 correlation coefficients, Z-scores and normalized ranks between AACE18 and the other 13 energy functions. Corresponding *p*-values are in columns 4-5, 7-8 and 10-11. *P*-values < 0.05 are in blue.

potential	*r*	95% confidence interval	*p*-value for *r*		Z-score	*p*-value for *Z*-score		rank	*p*-value for rank	
			2-sided	1-sided		2-sided	1-sided		2-sided	1-sided
1	2	3	4	5	6	7	8	9	10	11
AACE18	0.606	(0.533;0.670)	-	-	1.09	-	-	0.809	-	-
GOAP	0.587	(0.511;0.654)	8.61E-02	4.31E-02	1.13	3.10E-01	8.45E-01	0.821	2.10E-01	8.95E-01
AACE167	0.585	(0.504;0.647)	7.99E-02	4.00E-02	1.08	9.82E-01	4.91E-01	0.808	9.26E-01	4.63E-01
DFIRE	0.562	(0.483;0.632)	3.62E-03	1.81E-03	0.90	1.28E-02	6.40E-03	0.796	7.46E-01	3.73E-01
dDFIRE	0.547	(0.468;0.617)	2.24E-03	1.12E-03	0.87	2.97E-03	1.49E-03	0.790	3.92E-01	1.96E-01
AACE20	0.540	(0.460;0.610)	5.01E-04	2.51E-04	0.88	1.43E-03	7.17E-04	0.755	2.44E-03	1.22E-03
RF-CB-SRS-OD	0.533	(0.451;0.606)	7.62E-06	3.81E-06	0.99	1.38E-01	6.88E-02	0.791	1.89E-01	9.44E-02
RRCE20	0.531	(0.450;0.603)	3.24E-05	1.62E-05	0.88	1.52E-03	7.58E-04	0.751	7.04E-04	3.52E-04
RW	0.524	(0.441;0.599)	1.05E-05	5.27E-06	0.83	1.62E-03	8.11E-04	0.769	1.64E-01	8.21E-02
RWplus	0.518	(0.434;0.594)	2.25E-06	1.12E-06	0.83	2.86E-03	1.43E-03	0.770	3.62E-01	1.81E-01
OPUS-PSP	0.515	(0.430;0.590)	3.68E-07	1.84E-07	1.04	6.79E-01	3.39E-01	0.796	6.99E-01	3.50E-01
DOPE	0.508	(0.422;0.584)	6.74E-09	3.37E-09	0.89	5.16E-02	2.58E-02	0.787	7.50E-01	3.75E-01
MJ3h	0.493	(0.407;0.571)	1.06E-10	5.31E-11	0.74	4.64E-07	2.32E-07	0.710	8.94E-08	4.47E-08
RF-HA-SRS	0.424	(0.331;0.510)	8.87E-21	4.44E-21	1.04	2.23E-02	1.12E-02	0.796	3.41E-02	1.71E-02

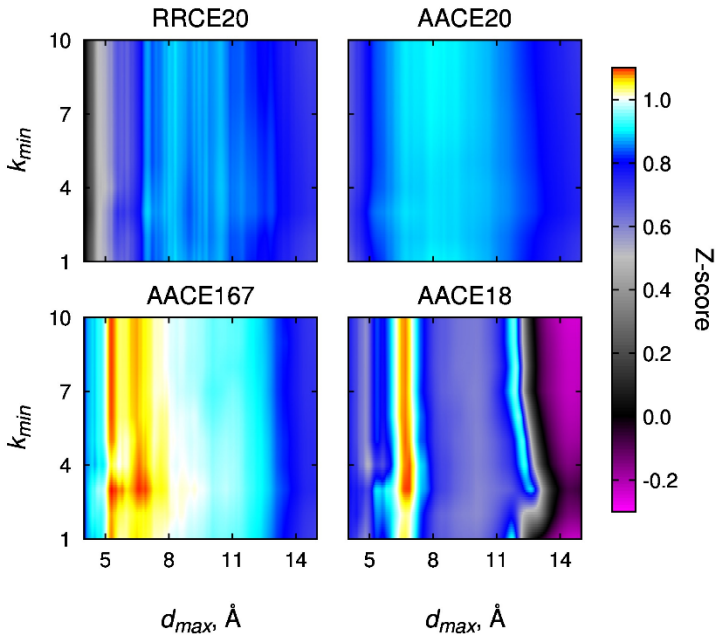

Figure S1: *Performance of the residue-residue and atom-atom contact potentials in best model recognition from CASP decoys*. The potentials derived at different values of sequence separation k_{min} and distance cut-off d_{max} were used to score models of 224 protein domains submitted to CASP rounds X and XI. The performance, measured as Z-score of the best model (the one with the highest GDT_TS score) averaged over all 224 evaluation units, is shown as heat map for RRCE20, AACE20, AACE167 and AACE18 potentials.

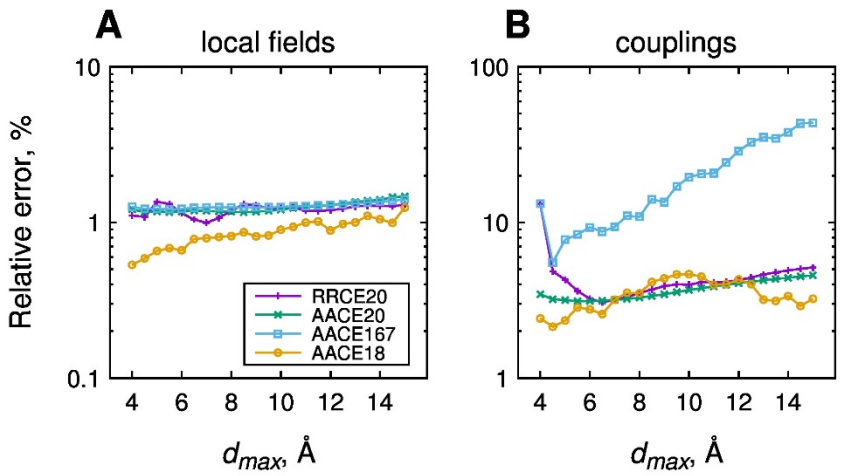

Figure S2: *Accuracy of the contact potentials parameters at different distance cut-offs.* The initial training set of 6,319 proteins was randomly split into halves. Each of the resulting subsets was used to train the statistical potentials at different distance cut-off d_{max} = 4 – 15Å with 0.5Å step, yielding in each case two sets of parameter estimates $\vec{\mathbf{h}}^{(1)}$, $\mathbf{J}^{(1)}$ and $\vec{\mathbf{h}}^{(2)}$, $\mathbf{J}^{(2)}$. Relative error was then calculated separately for (A) local fields $\vec{\mathbf{h}}$ and (B) couplings $\mathbf{J}$ using equation $\delta_{relative} = \left\|\vec{r}^{(1)} - \vec{r}^{(2)}\right\| \Big/ \left\|\vec{r}^{(1)} + \vec{r}^{(2)}\right\|$, where $\|\cdot\|$ is the l_2 vector norm. In the case of local fields, vector $\vec{r}$ is identical to vector $\vec{\mathbf{h}}$. For the coupling constants, $\vec{r}$ is composed of the upper triangle of matrix $\mathbf{J}$ plus the diagonal elements ($\mathbf{J}$ is symmetric, so the lower triangle was omitted). Relative errors $\delta_{relative}$ were calculated for five different random splits of the initial training set, and only the average values are shown on the plots.

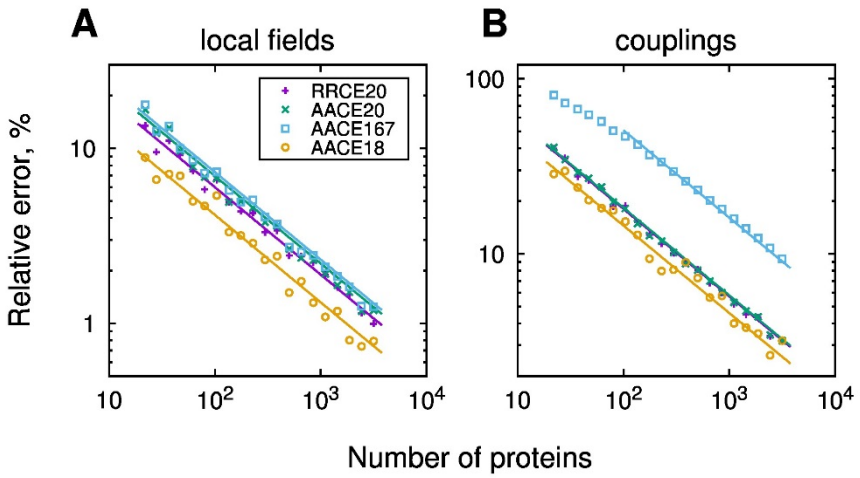

Figure S3: *Accuracy of the contact potentials parameters with varying sizes of the training set.* Using the procedure described in Figure S1, relative errors $\delta_{relative}$ for (A) local fields and (B) coupling constants were calculated for the randomly selected training subsets of different sizes ranging from 22 to 3159. The computed errors were fit by an empirically matched dependence $\delta_{relative} \sim 1/\sqrt{N}$, where N is the number of proteins used for training. Slight deviation of the AACE167 potential from this dependence (blue squares on the right-hand panel) is potentially caused by a very large number of parameters (~15,000), so that the system of equations (8) is underdetermined at small training set sizes N.

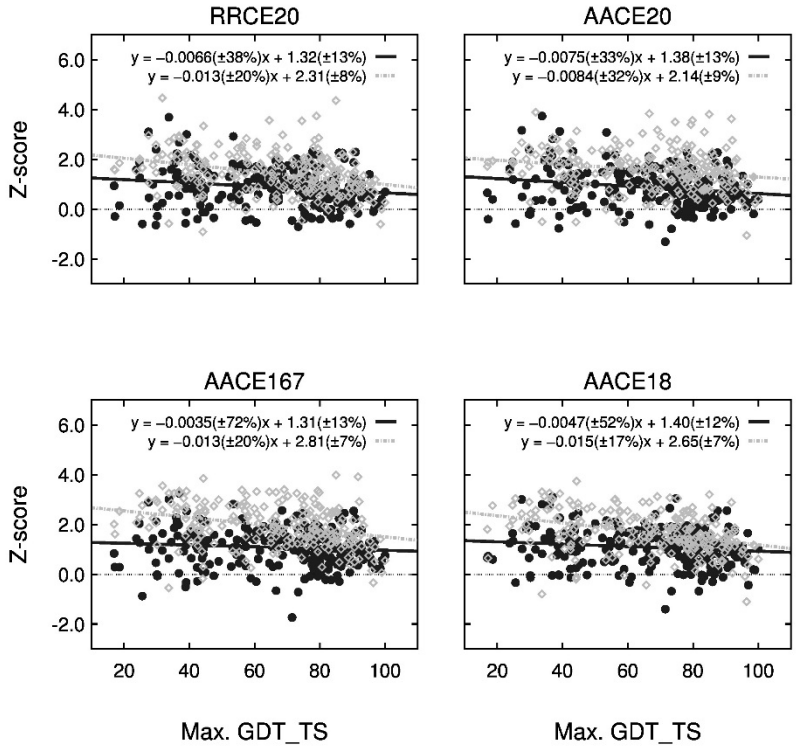

Figure S4: *Z-scores of the native structure (gray) and the highest accuracy model (black) in the CASP decoys depending on the decoys quality.* The GDT_TS score of the highest accuracy model (the best model according to CASP) was used as the measure of the decoys quality. For each of the 224 CASP decoy sets, the energy was calculated by the four contact potentials (see Methods in the main text),

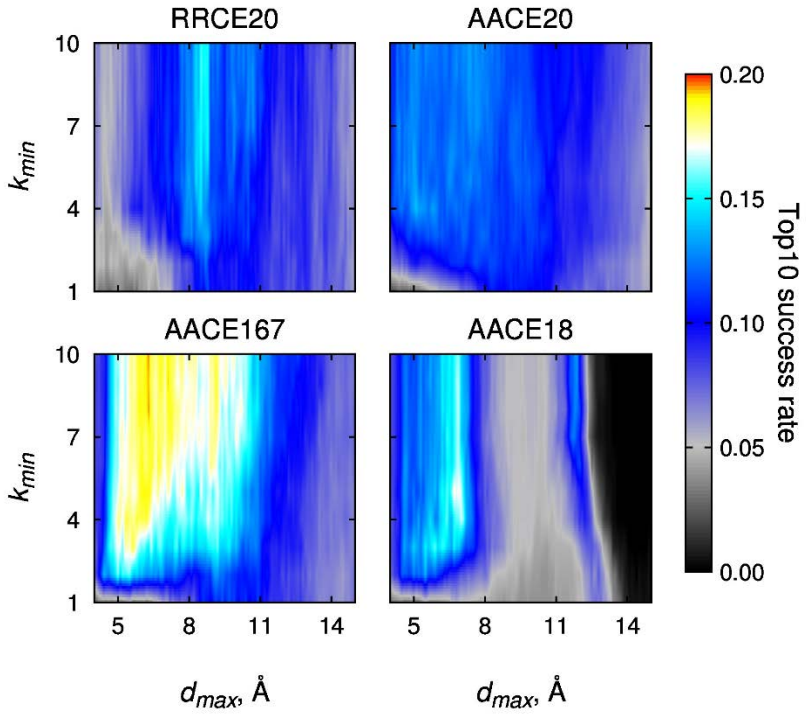

Figure S5: *Performance of the residue-residue and atom-atom contact potentials in near-native complex discrimination from low-resolution docking decoys*. Statistical potentials derived at different values of sequence separation k_{min} and distance cut-off d_{max} were used to score 100,000 unclustered matches for each of the 394 protein-protein complexes from DOCKGROUND Benchmark 4.0. Performance is measured in terms of the top-10 docking success rate (the fraction of complexes that have at least one near-native solution - acceptable or better quality according to CAPRI - among 10 best-scored models).

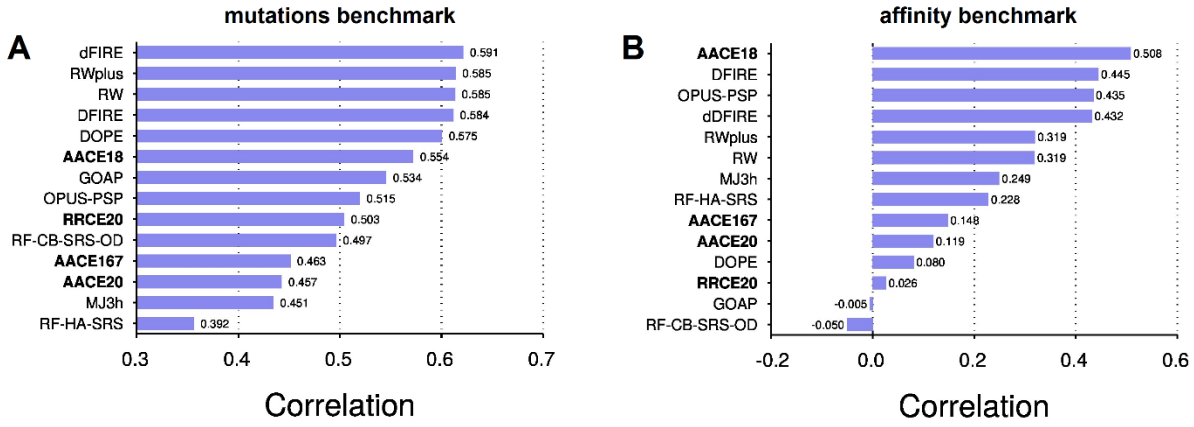

Figure S6: *Correlation of experimentally determined and calculated free energies.* (A) Pearson's correlation coefficient *r* between experimentally measured ($\Delta\Delta G_{\text{exp}}$) and calculated ($\Delta\Delta G_{calc}$) changes in folding free energies caused by point mutations over a set of 2,684 mutations for different knowledge-based energy functions. (B) The same scoring functions tested on their ability to recapitulate experimentally measured binding free energies (ΔG_{exp}) of 92 rigid-body complexes from Affinity Benchmark 2.0. The plot shows correlation coefficient *r* between ΔG_{exp} and ΔG_{calc} (see Methods). The RRCE20, AACE20, AACE167 and AACE18 potentials were derived at d_{max} = 8.0 Å and k_{min} = 3. Scoring functions on both panels are sorted by their performance according to *r*.

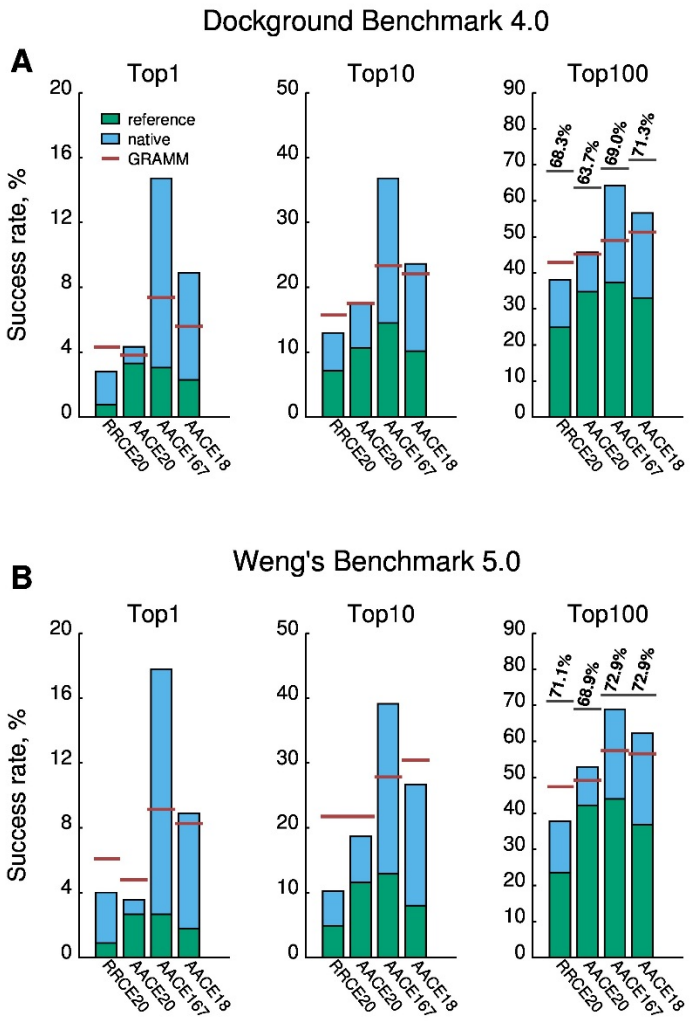

Figure S7: *Ranking of the native and reference structures in low-resolution docking decoys*. After scoring and clustering of top 100,000 matches from GRAMM (see Methods and caption to Fig. 6 in the main text for details), we checked whether the native (bound conformation, blue bars) and reference (unbound superimposed onto bound, green bars) is scored higher than any of the top 1,10 and 100 docking clusters. The fraction of such cases is plotted for (A) DOCKGROUND Benchmark 4 and (B) Weng's Benchmark 5. For comparison, docking success rates from Fig. 6 are shown by horizontal red lines. The top100 plots also show the maximal achievable docking success rates: black lines show the fraction of cases for which at least one docking cluster is of acceptable or better quality (see Table 1), regardless of its score.